Real-Time Edge AI for Distributed Systems (READS)

Disentangling Beam Losses in the Fermilab Main Injector Enclosure Using Real-time Edge AI

Kyle J. Hazelwood
ICFA ML 2022
November 02, 2022

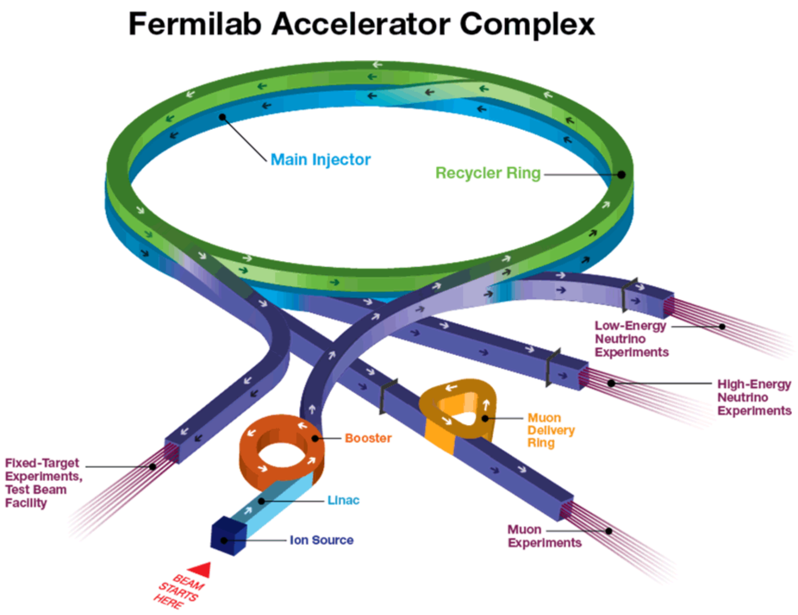
In partnership with:

NORTHWESTERN UNIVERSITY

Fermilab Main Injector and Recycler

- Main Injector
  - 8-120 GeV (150 GeV) synchrotron proton accelerator
  - 3.3 kM (2.05 mile) machine circumference
  - Delivers 120 GeV, 1MW beams to NuMI beamline experiments
  - Delivers 120 GeV resonant extracted beams to Switchyard experiment beamlines
- Recycler
  - 8 GeV permanent magnet ring
  - 3.3 kM (2.05 mile) machine circumference
  - Originally purposed as an antiproton storage ring for TeVatron collider operations
  - Now used as a proton stacker for high intensity NuMI beams (injecting to Main Injector)
  - Accumulates and bunches beam for g-2 experiment

Both machines reside in the same tunnel!



Project Overview

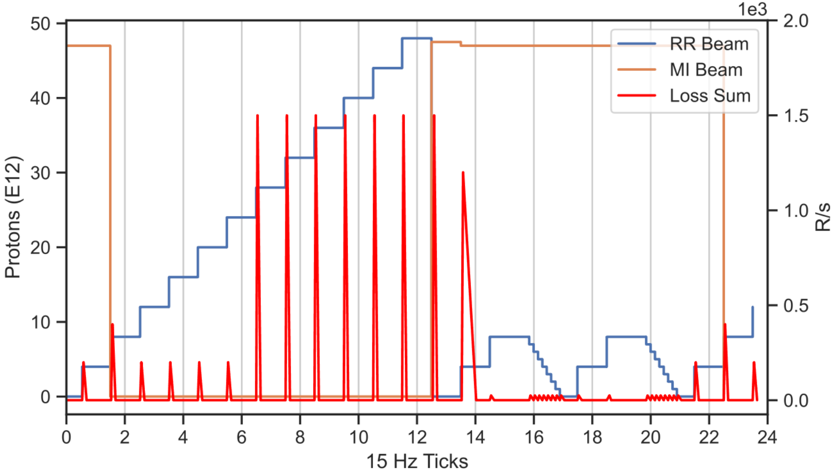
- Main Injector and Recycler share an enclosure
- Both machines can and do often have high intensity beam in them simultaneous
- Both machines can generate significant beam loss
- The machine origin of a beam loss is often hard to distinguish
- Often losses from one machine end up tripping the machine permit of the other resulting in unnecessary beam downtime

The projects aims to deploy a machine learning model on a FPGA that when fed streamed beam loss readings from around the Main Injector complex, will infer in real-time the machine loss origin

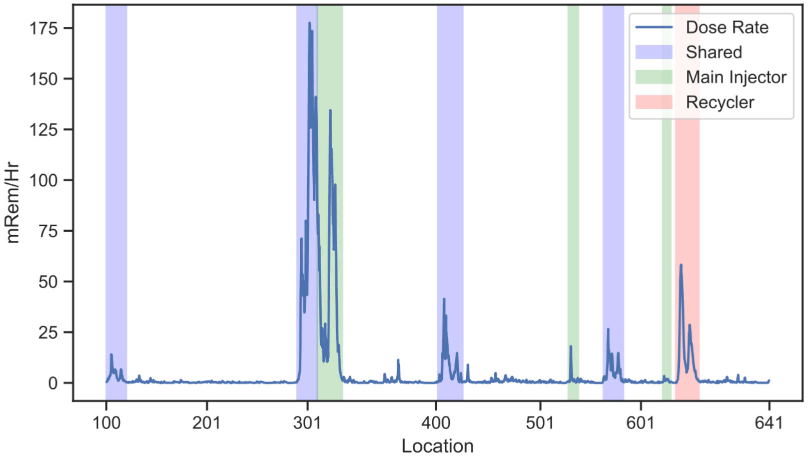
Main Injector tunnel
Recycler (top) Main Injector (bottom)

Project Overview

- Using time, location and state of the machine, machine experts can sometimes attribute loss to a particular machine
  - This suggests a Machine Learning (ML) model may be trainable to automatically attribute loss and replicate or improve upon the expert's ability



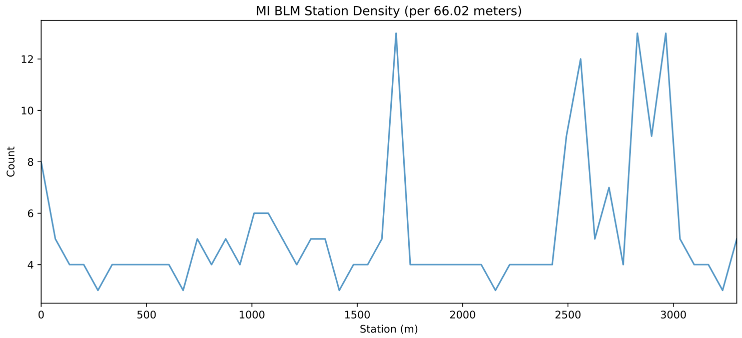
Example illustration of overlapping beam events and losses in the MI and RR accelerators



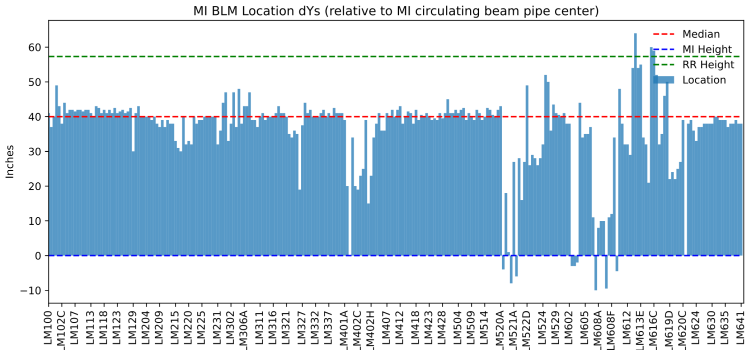
Location dependency of MI and RR beam loss as seen from tunnel activation residual doses

Beam Loss Monitors (BLM)

- Glass Ionization Chambers
- 259+ BLMs, 7 BLM nodes distributed around the MI complex
- BLM nodes provide ACNET loss readings
  - **Hardware unable to stream all BLM readings simultaneously at fastest readout frequency**
- BLM nodes tied into the machine permit
- Recorded the location of all BLM in the Main Injector tunnel
- Assume BLM locations will have to be controlled once a ML model has been trained, else risk having to re-train
- 23% of the BLMs exist in ~10% of tunnel



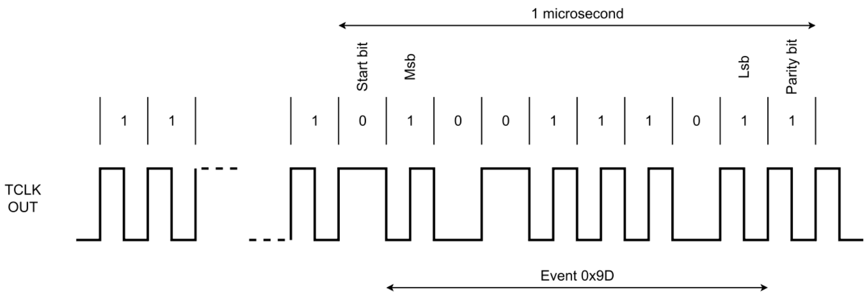
Beam Loss Monitor



NORTHWESTERN UNIVERSITY
Fermilab

TCLK and MDAT

- TCLK (TeVatron Clock)
  - 10 MHz Event Clock
  - Provides current machine program and status
  - Originates from Timeline Generator (TLG) hardware
- MDAT (Machine Data)
  - 720 Hz
  - Originates from various hardware around Main Injector Complex
    - Low Level RF
    - Beam Current Monitor Front Ends
    - Main Injector Ramp Regulation Front End (MECAR)

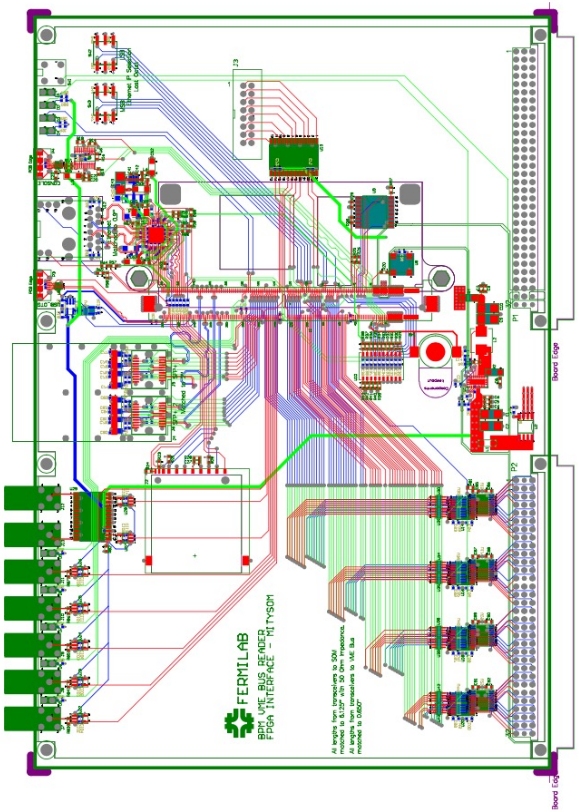


TCLK (10 MHz Clock)

VME Bus Reader (Pirate) Cards

- One of the requirements of the project was not to interfere in any way with the normal operation of the existing BLM system
  - The purpose of the Pirate cards are to monitor BLM digitizer polls and transmit the data over ethernet

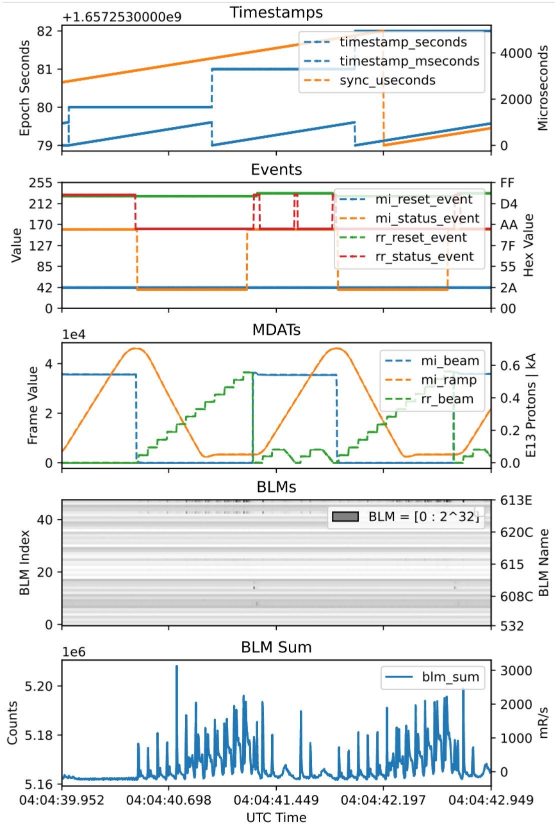
- Pirate cards are MitySOM Cyclone 5 FPGAs on custom VME carrier boards
- Each BLM node has a Pirate card installed in it (7 cards total)
- All BLM channels are available to stream simultaneously
- Also transmits TCLK and MDAT readings
- TCLK and MDAT are monitored to provide microsecond timestamping
  - Used to synchronize streams across multiple cards
- Data is streamed via UDP in DDCP protocol format
- Data frequency is 333 Hz (current rate of digitizer polling)
- Each card streams 0.4-0.6 Mb/s (dependent on number of BLM channels in crate)



VME Bus Reader (Pirate) card

Datasets

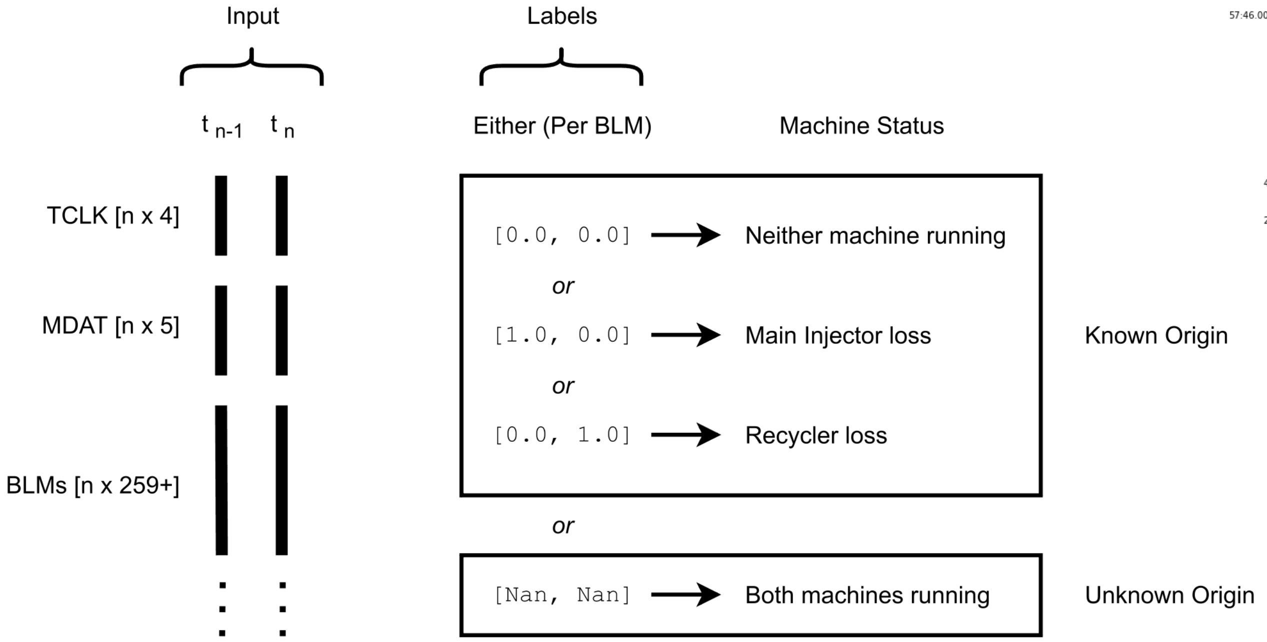
- Sample Dataset
  - 15/33 Hz
  - Data taken from machine operations via ACNET
  - Includes all 259 operational BLMs, TCLK, and MDAT data
  - Taken throughout the 2020/2021/2022 runs
- High Frequency Dataset
  - 333 Hz (BLM node digitizer poll rate)
  - Data from VME Bus Reader (Pirate) cards commissioned June 2022
  - Same data as Sample datasets albeit faster
- Study Datasets
  - 33/333 Hz
  - Data taken from 2021/2022 dedicated end of run studies
  - Includes all the same data as the Sample and High Frequency datasets
  - Timeline altered so that only Main Injector or Recycler had beam at any time
  - Beam losses purposefully generated in both machines using various machine miss-configurations to not bias a model towards standard running
  - All beam loss attributable to a machine



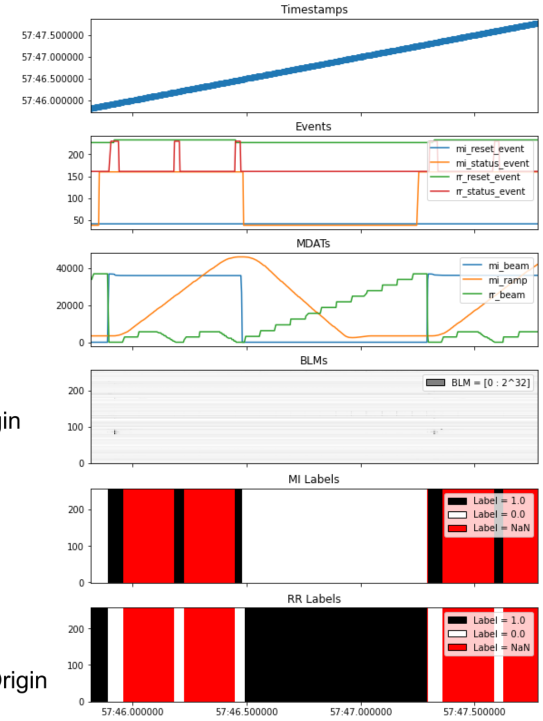
Few second example of Pirate card stream

Data Labeling

- Data labeling automated using MDAT, and TCLK information



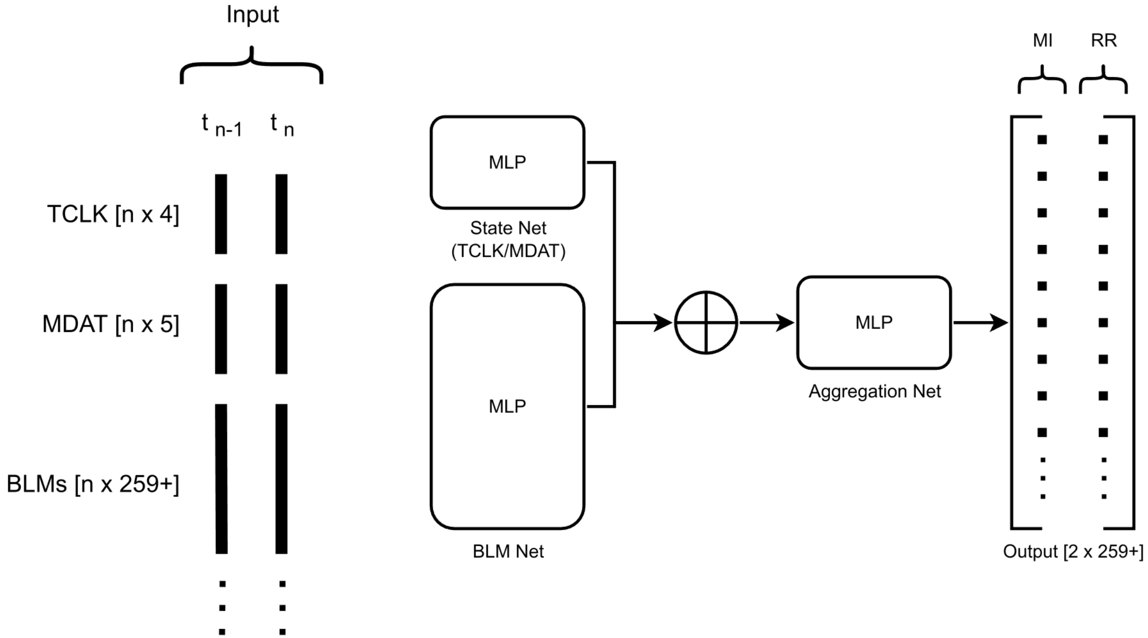
Labeling scheme



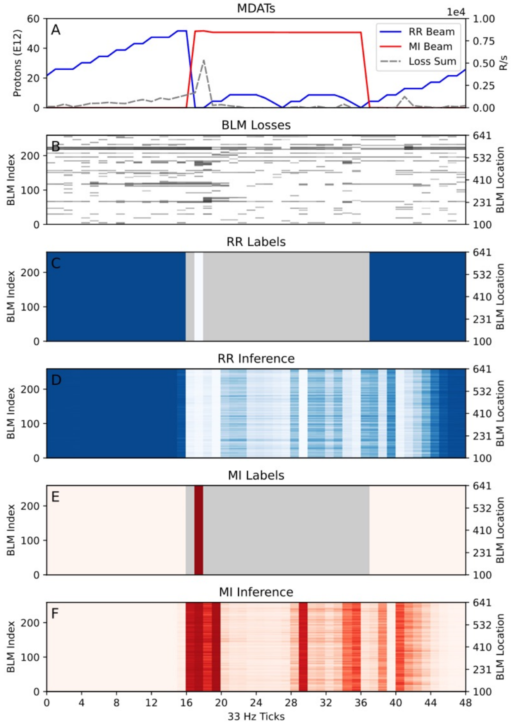
Data labeling example

ML Model Architecture: Phase 1, Data-Type-Specific Aggregation (DBLN)

Objective: Assign BLM-wise probabilities for that loss originating in MI/RR



Model architecture



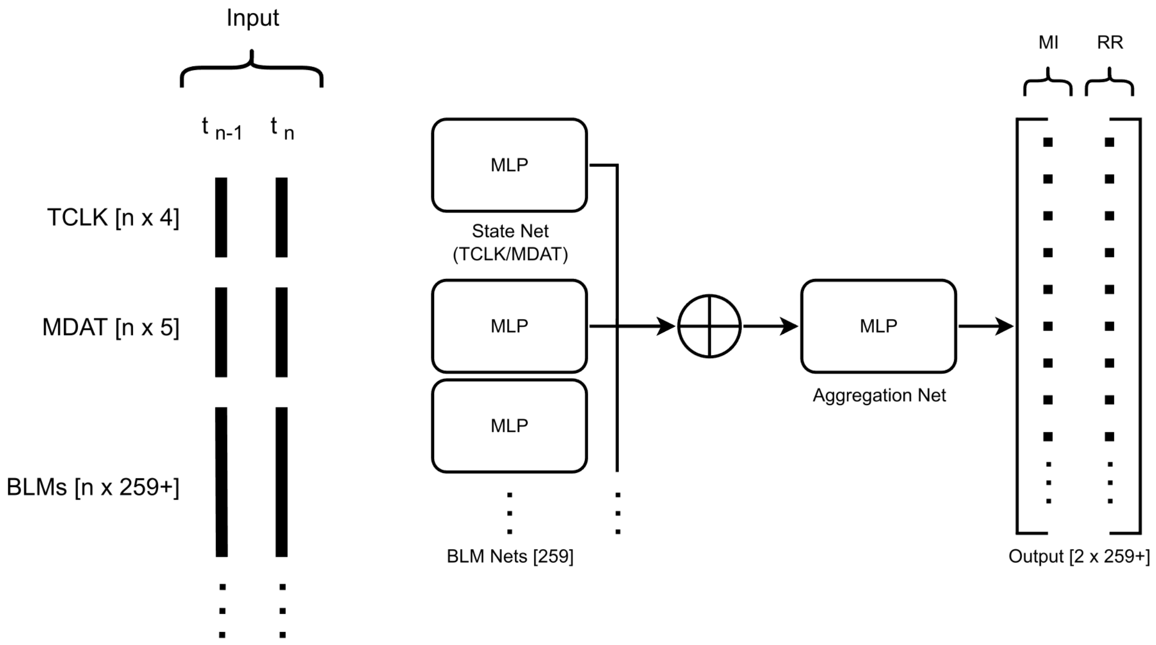
Model inference

NORTHWESTERN UNIVERSITY

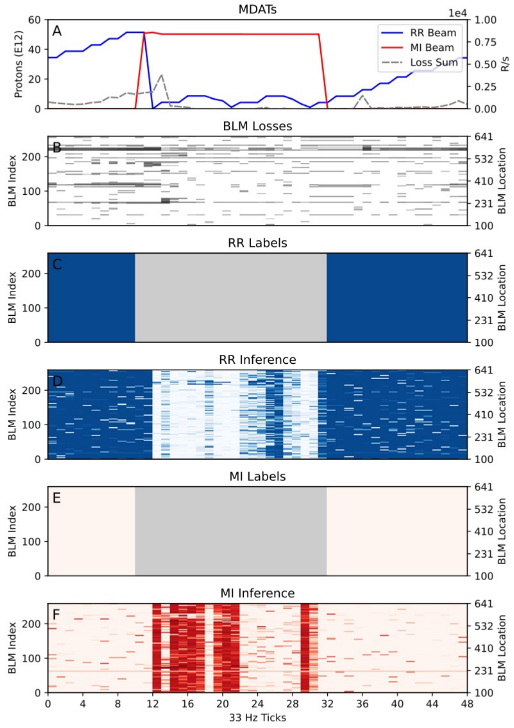
Fermilab

ML Model Architecture: Phase 2, Forcing Locality (ManyModels)

Objective: Assign BLM-wise probabilities for that loss originating in MI/RR



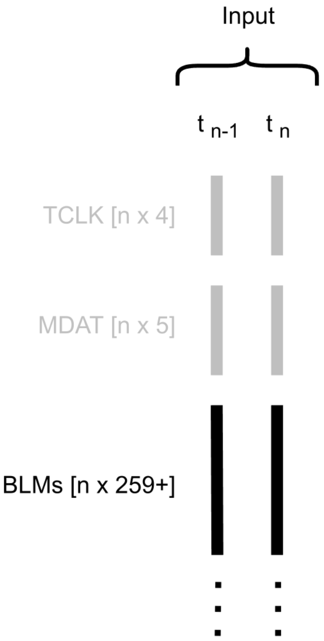
Model architecture

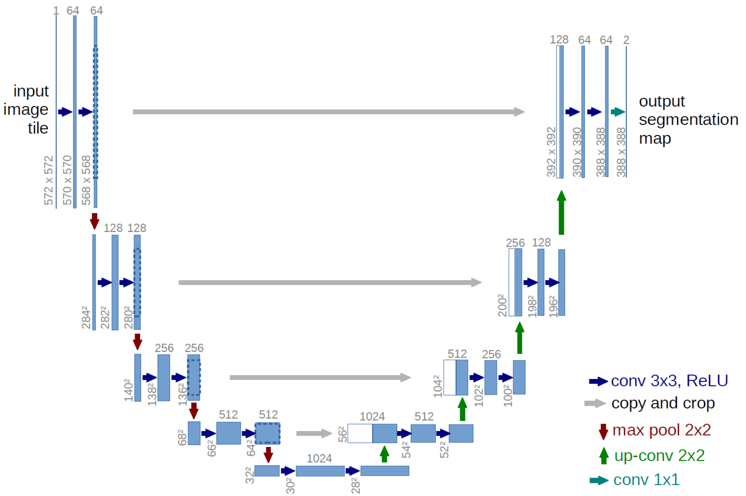


Model inference

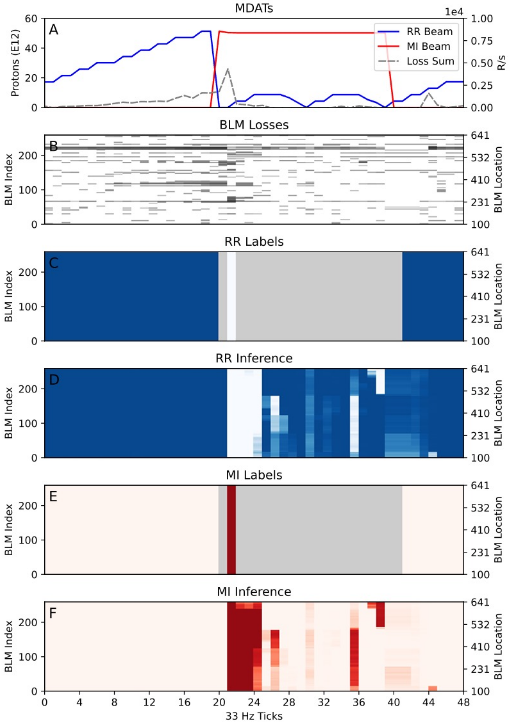
ML Model Architecture: Phase 3, Varying Receptive Fields (UNet)

Objective: Assign BLM-wise probabilities for that loss originating in MI/RR



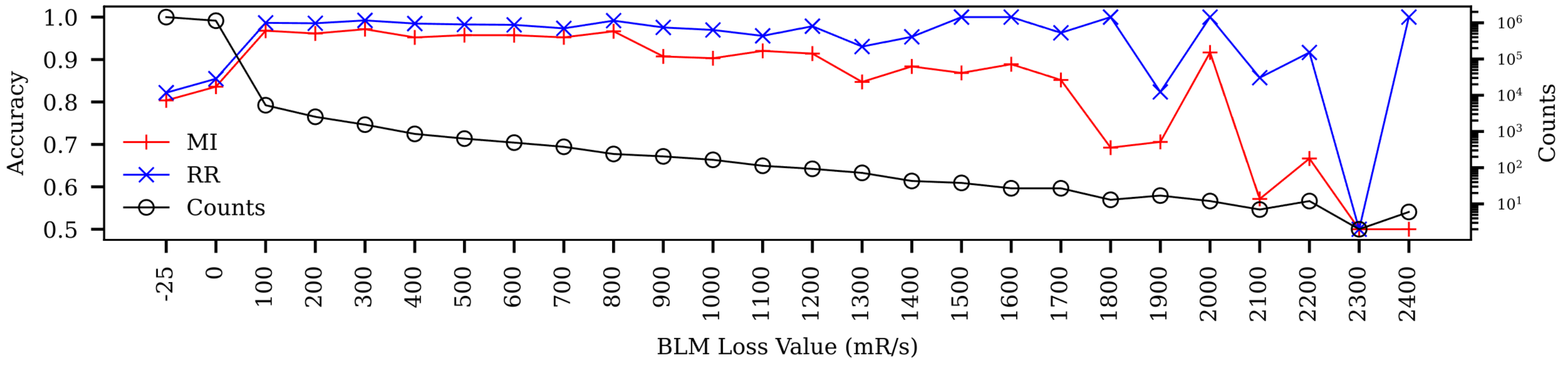


Model architecture (example)



Model inference

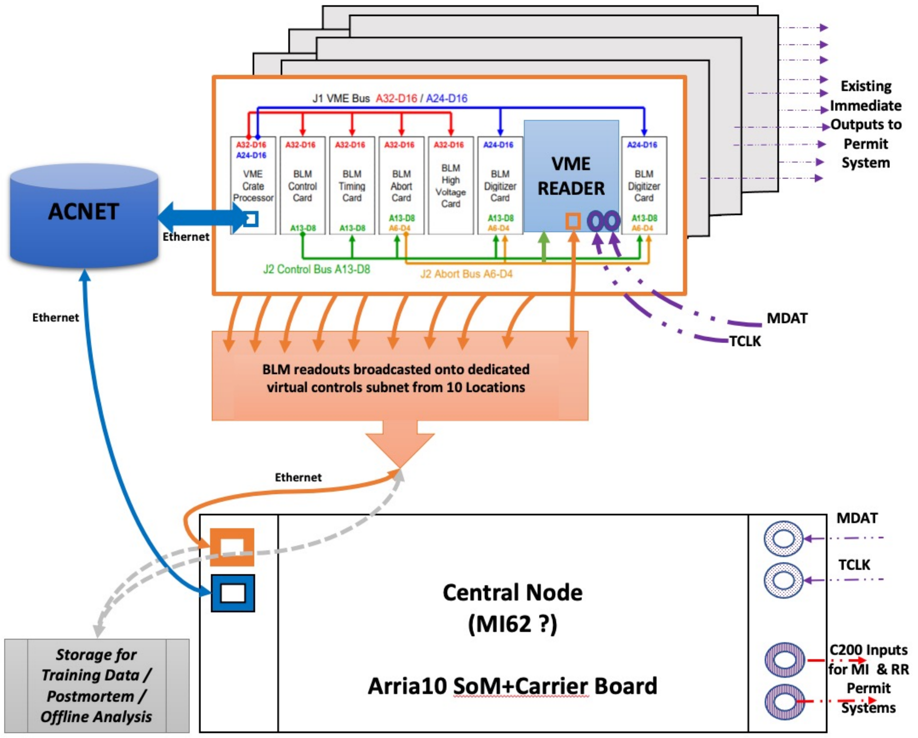
ML Model Architecture: Phase 3, Varying Receptive Fields (UNet)



Example inference accuracies for beam loss values of interest

Central Node

- Central node is an Aria10 FPGA SOM
- ML model will be deployed on FPGA
- Two HPS Arm cores and ethernet pots
  - One dedicated to ingesting VME bus reader card streams
  - One dedicated to and EPICS IOC to provide control system readings and waveforms
- Has inputs for MDAT and TCLK
- Has TTL outputs intended for MI and RR c200 permit input



Central node data paths

Remaining Schedule / Work

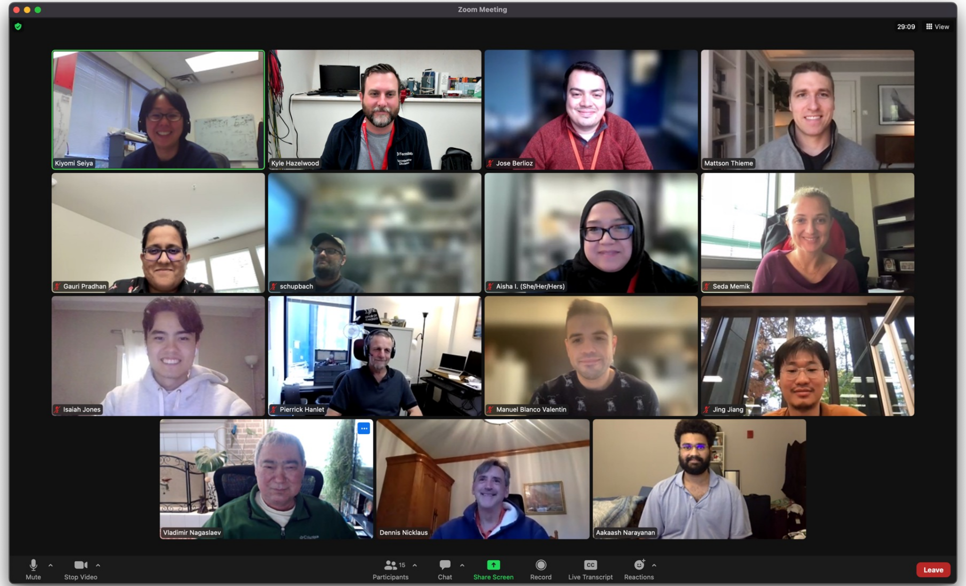
- Build and compile ML model for Aria10 Quartus project using hls4ml
- Build out advanced readouts from EPICS IOC
- Explore even lower latency VME Bus Reader card stream transmissions
- Optimize model(s) further
  - Tune hyperparameters
  - Prune/adjust architecture
  - Explore further data pruning,
  - Compare normalization and standardization methods
- Investigate model robustness
  - BLM readings missing or incorrect
  - TCLK, MDAT jitter
- Investigate parameter quantization
- Investigate active learning
- Investigate loss prediction
- Research figure of merit to compare expert loss attribution to model inference
- Tie Central Node into Main Injector and Recycler machine protection systems

NORTHWESTERN UNIVERSITY Fermilab

Summary

- VME Bus Reader (Pirate) cards were built to stream BLM readings from the enclosure
- Large amounts of data have been collected and continue to be collected for model training
- A promising ML model has been created to disentangle beam loss in the Main Injector enclosure
- Work is underway to implement our model on a Central Node FPGA
- Work continues to further optimize the model and explore its robustness and durability

Acknowledgments



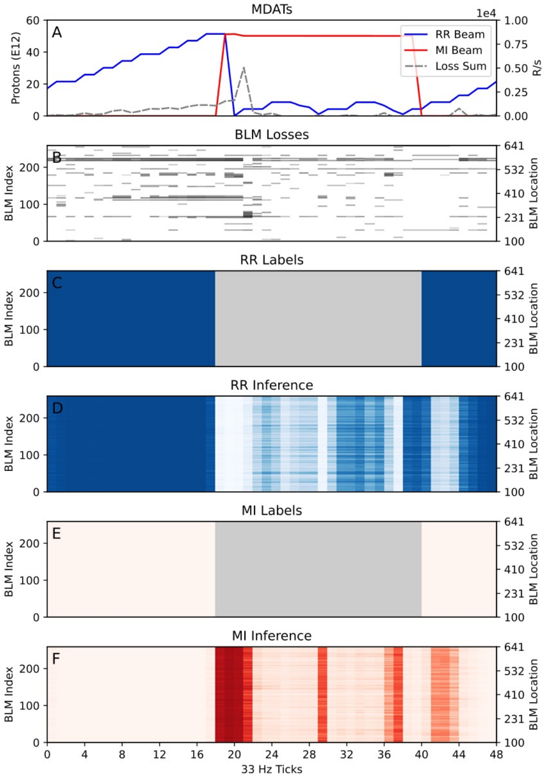
Fermilab
M.R Austin, J.M. Arnold, K.J. Hazelwood, P. Hanlet, M.A. Ibrahim, A. Narayanan, D. J. Nicklaus, G. Pradhan, A.L. Saewert, B.A. Schupbach, K. Seiya, R.M. Thurman-Keup, N.V. Tran

Northwestern University
J. Jang, H. Liu, S. Memik, R. Shi, M. Thieme, M. Valentin

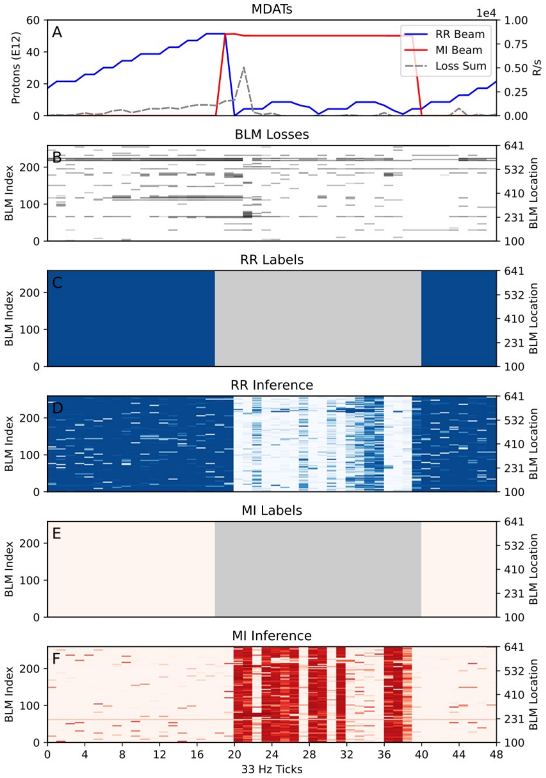
Thank You

- K. Seiya *et al*, "Accelerator Real-time Edge AI for Distributed Systems (READS) Proposal" (March 2020) https://arxiv.org/abs/2103.03928
- K.J. Hazelwood *et al*, "Real-Time Edge AI for Distributed Systems (READS): Progress on Beam Loss De-Blending for the Fermilab Main Injector and Recycler" (August 2021) https://lss.fnal.gov/archive/2021/conf/fermilab-conf-21-603-ad-scd.pdf
- J. Berlioz *et al, "*Synchronous High-Frequency Distributed Readout for Edge Processing at the Fermilab Main Injector and Recycler" (August 2022) https://napac2022.vrws.de/papers/mopa15.pdf
- M. Thieme *et al*, "Semantic Regression for Disentangling Beam Losses in the Fermilab Main Injector and Recycler" (August 2022) https://napac2022.vrws.de/papers/mopa28.pdf

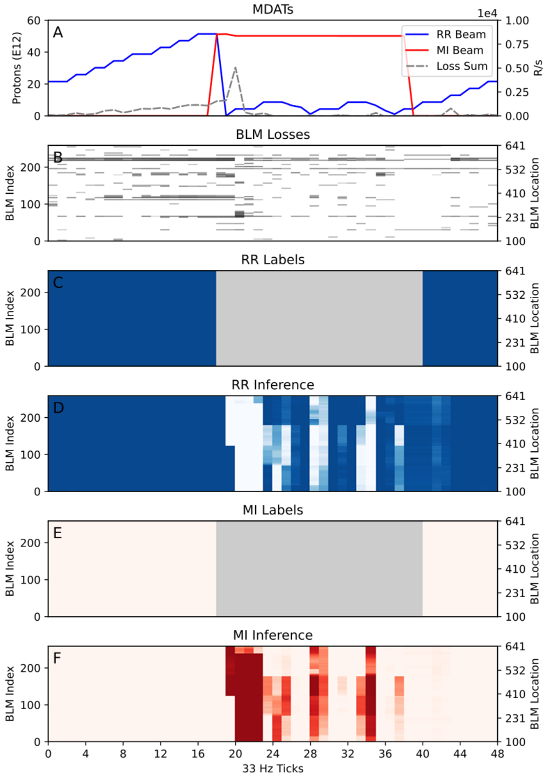
Model Comparison



DBLN



Many Models

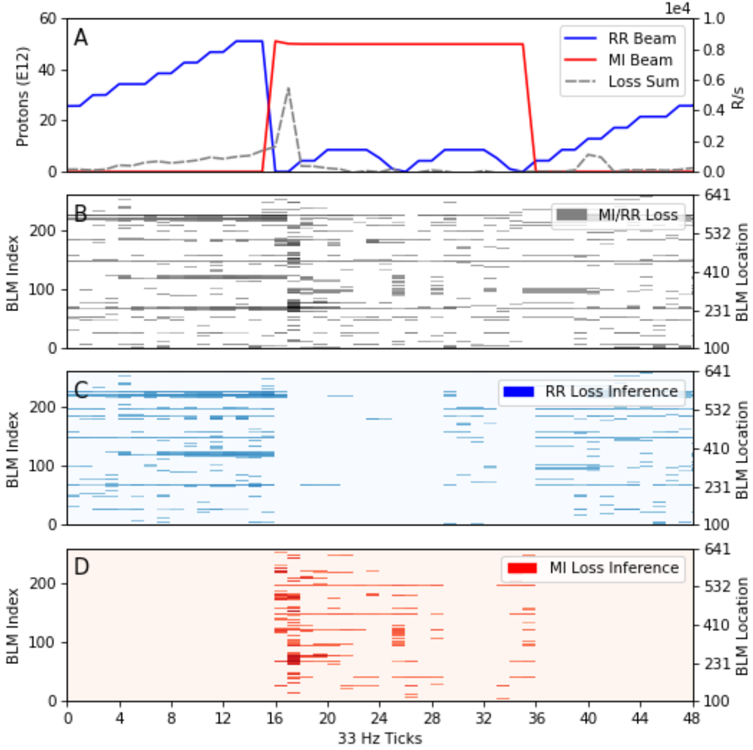


UNet

NORTHWESTERN UNIVERSITY

Fermilab

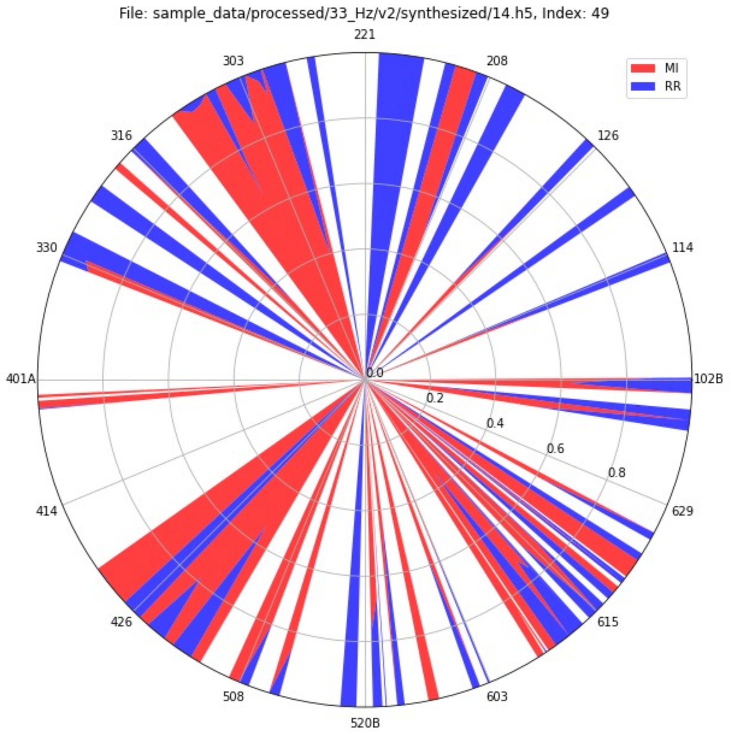
ML Model Inference (continued)



Example model inferred losses

Datasets (continued)

- Synthesized Dataset
  - 33 Hz
  - Using Sample and Study Datasets
  - Use known losses (attributed to one machine) and sum with known losses attributable to the other machine
  - Resulting labels are percentages of loss per BLM attributed to a machine
  - Will be used to attempt a semi-supervised model training



Example of synthesized data labeling